PSTAT 5LS Lab 6

Professor Miller

Week of May 22, 2023

Section 1

Learning Objectives

R Learning Objectives

1. Interpret R output providing confidence intervals and hypothesis tests for inference on population means.
2. Create a graphical display of the t distribution.

Statistical Learning Objectives

1. Learn about the t distribution
2. Get experience making confidence intervals for population means
3. Understand hypothesis tests for population means

Functions Covered in this Lab

1. `pt()`
2. `qt()`
3. `plot_t()`
4. `t.test()`

Section 2

Lab Tutorial

One Population Mean

This week, we are shifting our focus from categorical data (proportions) to *numeric* data (means). Inference for one mean is a new tool in our statistical toolkit that will let us answer different questions about data. Keep in mind that the parameter we're interested in now is μ (mu), the population mean.

The t Distribution

The t **distribution**, like the standard normal distribution ($N(0, 1)$) that we've seen before, is symmetric about zero and bell-shaped. The t distribution, however, has "heavier tails" than the normal distribution. This is so that we can capture the increased uncertainty introduced when we use the sample standard deviation s to estimate the population standard deviation σ.

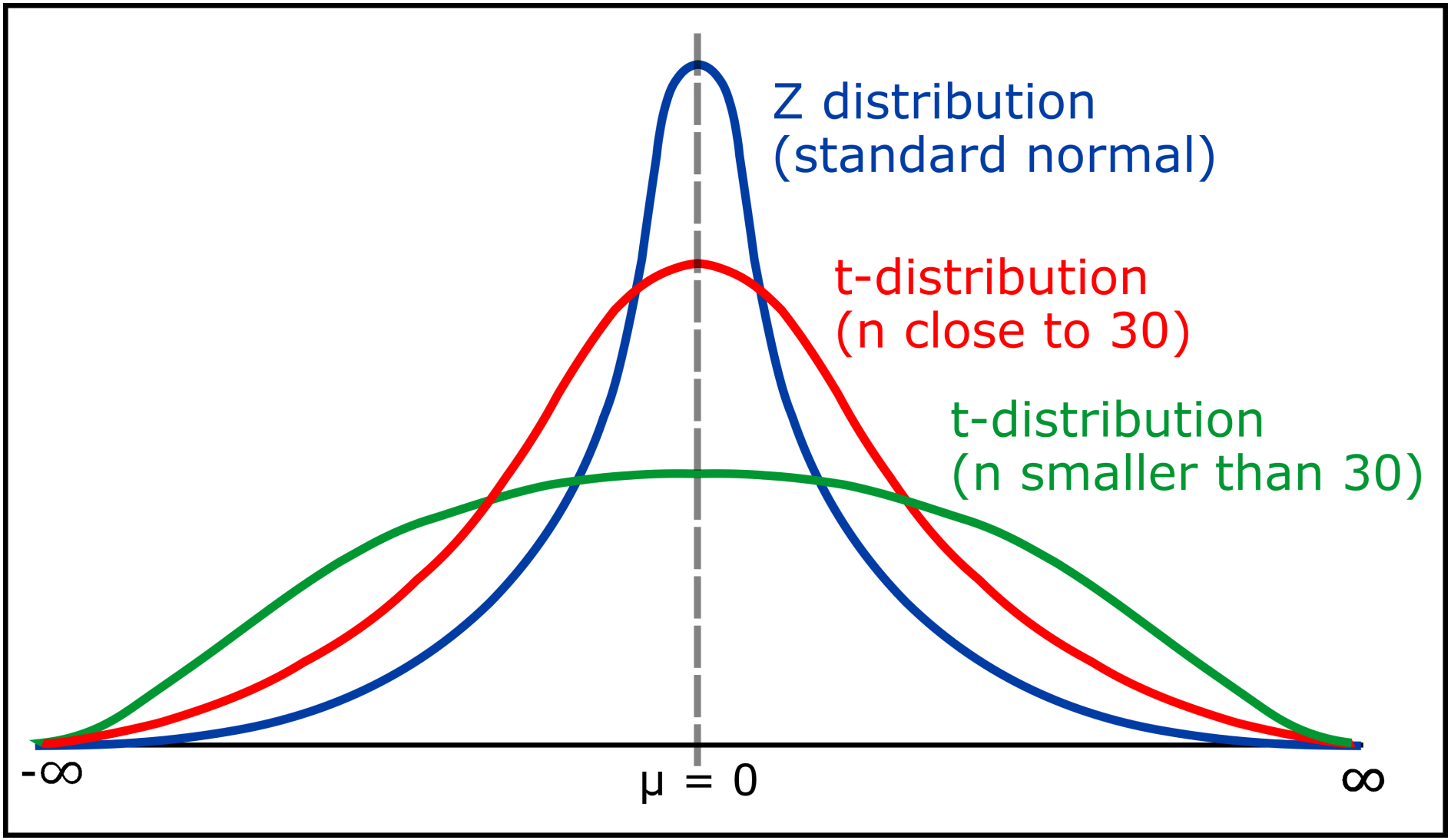
The t Distribution

The t distribution is a "family" of distributions, meaning there are an infinite number of t distributions. The same is true of the normal distribution – there are infinite different normal distributions.

We describe which normal distribution we're talking about with two parameters, the mean μ and the standard deviation σ.

For t distributions, we identify which one we're talking about using just one parameter, the **degrees of freedom (df)**. As the degrees of freedom increase, the t distribution gets closer and closer to a standard normal distribution, $N(0, 1)$. Check this out:

t Distribution Versus the Standard Normal Distribution



Finding Probabilites Under a t Distribution

We can find probabilities related to the t distribution using the `pt()` function. The `pt()` function is analogous to the `pnorm()` function we used with normal distributions.

Arguments to send to `pt()`:

- `q`: The observation is sometimes called a **quantile**, thus the **q** notation.
- `df`: This is the degrees of freedom for the sample; this week, we can find `df` by computing $n - 1$, where n is the sample size.
- `lower.tail`: By default, this argument is set to `TRUE`, meaning that we *want* the lower tail (i.e., to shade to the left). If we *don't want* the lower tail, and we actually want the *upper* tail (i.e., to shade to the right), we should set this to `FALSE`.

The mean of a t distribution is 0, and the standard deviation of a t distribution is a function of the the degrees of freedom is 1, so we *don't* need to specify the mean and the standard deviation to use `pt()`.

Graphing the t Distribution

We can also use the `plot_t()` function in the `stats250sbi` package to make a graphical display of the t distribution. As you might expect, `plot_t()` is similar to `plot_norm()`.

Arguments to send to `plot_t()`:

- `df`: This is the degrees of freedom for the sample; this week, we can find `df` by computing $n - 1$, where n is the sample size.
- `shadeValues`: This is the value(s) that we wish to identify, and shade either to the left or to the right of. To include two values we will combine them by using the `c()` function.
- `direction`: This is the direction to shade. The choices are: "less", "greater", "beyond", "between". The text must be written in double quotes.
- `col.shade`: Optional color choice for the graph.

Finding the Quantile on a t Distribution

Also, we can use `qt()` to get quantiles of the t distribution. The `qt()` function will be helpful to find t^* critical values needed for confidence intervals for means.

Arguments to send to `qt()`:

- `p`: The probability to the **left by default** of the quantile we wish to find. If we want the probability to the *right*, we should tinker with `lower.tail` as specified below.
- `df`: This is the degrees of freedom for the sample; this week, we can find `df` by computing $n - 1$, where n is the sample size.
- `lower.tail`: By default, this argument is set to `TRUE`, meaning that we *want* the lower tail (i.e., to shade to the left). If we *don't want* the lower tail, and we actually want the *upper* tail (i.e., to shade to the right), we should set this to `FALSE`.

Back to the Penguins!

Let's say we want to construct a confidence interval for the population mean flipper length of penguins living in the Palmer Archipelago, or conduct a hypothesis test for that mean.

In order to use our machinery for constructing confidence intervals and performing hypothesis tests for means, we need two conditions to hold. **What are they?**

Conditions for a CI or HT for the Population Mean

The conditions are

1. The observations must be *independent* of one another.
2. When the sample is small, we require that the *sample observations come from a normally distributed population*. We can relax this condition more and more for larger and larger sample sizes.

How can we check that we meet both of these conditions?

Checking the Conditions for a CI or HT for the Population Mean

To check the conditions:

1. Verify that we have taken a random sample from the population. If we don't have a random sample, we should consider whether it's reasonable or not to believe that the observations are independent of one another.
2. Examine a histogram of the sample data. When the sample is small, we require that the sample observations come from a normally distributed population. We can relax this condition more and more for larger and larger sample sizes.

- If $n < 30$, we will verify that the distribution is approximately bell-shaped. Slight skew is okay.
- If $n \geq 30$, we will verify that the distribution is approximately bell-shaped. If it is not and if there are no extreme outliers, then the sampling distribution of sample means will be approximately normal under the Central Limit Theorem.

Is the Penguins Data Observations Independent of One Another?

Do you think that the penguins data is from a random sample? If not, do you think that the observations of penguins are independent of one another?

Is the Penguins Data Observations Independent of One Another?

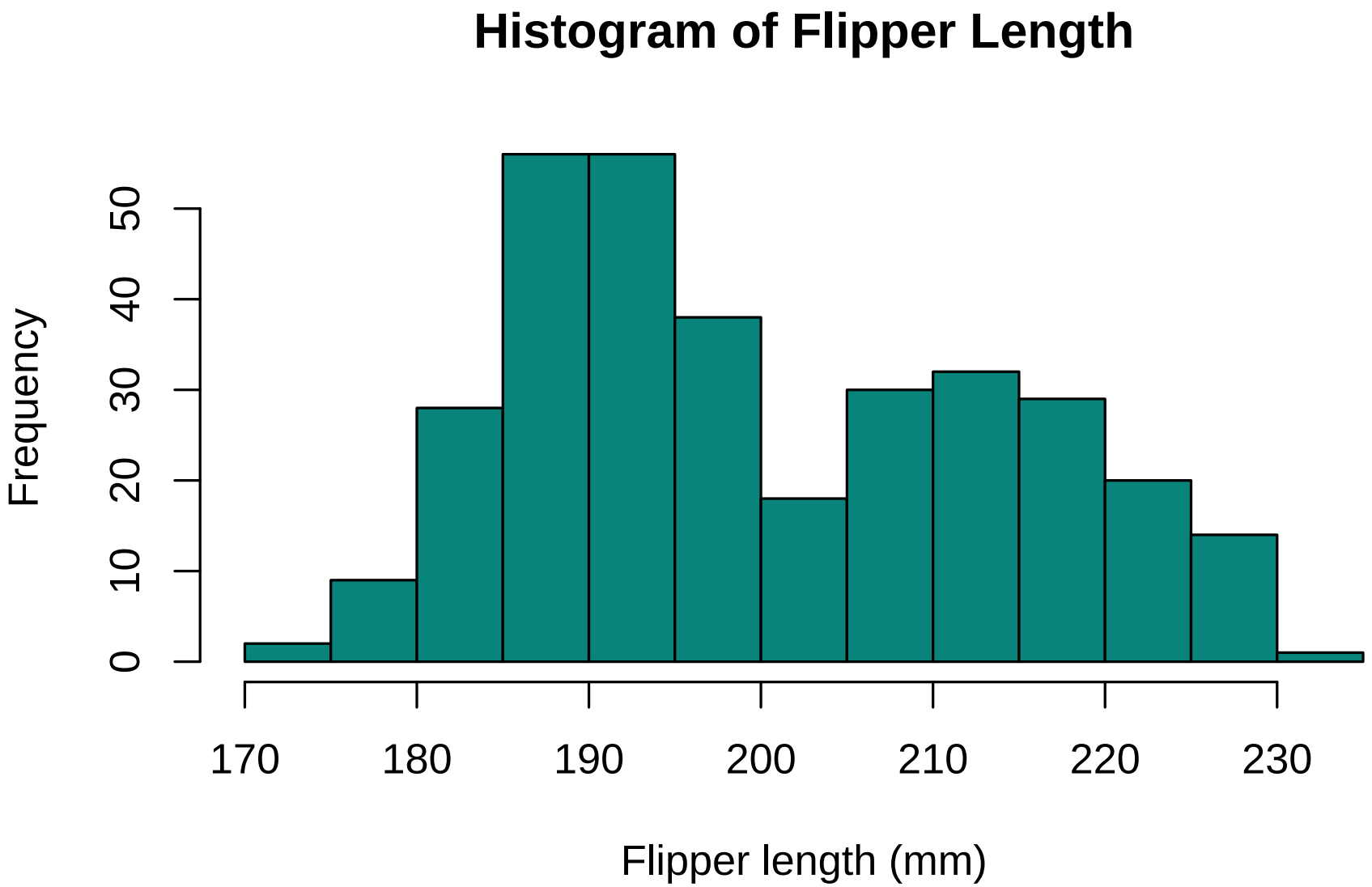
We are not told if this data is from a random sample. It might be reasonable to assume that each penguin's information was collected independently of another penguin, so we can proceed.

Is the Histogram of Sampled Penguin Flipper Length Approx. Normal?

Let's make a histogram in the `tryit1` code chunk of your notes document of the penguin flipper length. Don't forget to first run the chunk to read in the data!

Do you think that the distribution is approximately bell-shaped?

Histogram of Sampled Penguin Flipper Length



Considering the Sample Size

Although we observed a histogram that does not appear to be approximately bell-shaped, we can relax the condition knowing that we have a sample size that is much larger than 30.

Creating a Confidence Interval for the Mean Flipper Length

Let's construct a 90% confidence interval for the population mean flipper length of penguins living in the Palmer Archipelago.

Just like `prop_test()`, we can give `t.test()` the relevant information and have R compute the confidence interval for the mean flipper length.

Try this code out in the `tryIt2` code chunk in your notes document.

```
t.test(penguins$flipper_length_mm,
       conf.level = 0.9)
```

```
##
##  One Sample t-test
##
## data:  penguins$flipper_length_mm
## t = 261.66, df = 332, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  199.7001 202.2338
## sample estimates:
## mean of x
##   200.967
```

Confidence Interval Interpretation for One Mean

We estimate with 90% confidence that the mean flipper length for all penguins in the Palmer Archipelago is between 199.7mm and 202.2mm.

Hypothesis Test for a Mean

We can also use `t.test()` to help us run a hypothesis test about a population mean.

Let's consider this scenario. Suppose that the average flipper length of penguins in general is known to be 199 mm. Is the average flipper length for the penguins living on Palmer Archipelago *larger than 199 mm*? Use $\alpha = 0.05$.

Thus we are testing:

$H_0 : \mu = 199$

$H_a : \mu > 199$

where μ is the mean flipper length in the population of penguins living in the Palmer Archipelago.

Using t.test() to Run this Hypothesis Test

Just like prop_test(), we can give t.test() the relevant information and have R run the hypothesis test.

Try this code out in the tryIt3 code chunk in your notes document.

```
t.test(penguins$flipper_length_mm,
       mu = 199,
       alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  penguins$flipper_length_mm
## t = 2.561, df = 332, p-value = 0.00544
## alternative hypothesis: true mean is greater than 199
## 95 percent confidence interval:
##  199.7001      Inf
## sample estimates:
## mean of x
##   200.967
```

Conclusion for the Hypothesis Test

Because our p-value of 0.00544 is less than the significance level of 0.05, we reject the null hypothesis.

Our analysis suggests that the mean flipper length for the population of penguins on the Palmer Archipelago is greater than 199 mm.